Discovery of Association Rules in Medical Data

Srinivas Doddi

Los Alamos National Laboratory, P. O. Box 1663, MS B265,

Los Alamos, NM 87545, USA. Email: srinu@c3.lanl.gov.

Achla Marathe (Contact author),

Los Alamos National Laboratory, P. O. Box 1663, MS B265,

Los Alamos, NM 87545, USA. Email: achla@c3.lanl.gov.

Phone: (505) 667-9034.

Fax: (505) 665-5638.

S. S. Ravi

Department of Computer Science,

University at Albany - State University of New York,

Albany, NY 12222, USA. Email: ravi@cs.albany.edu.

This author's research was supported by NSF Grant CCR-97-34936.

David C. Torney

Los Alamos National Laboratory, P. O. Box 1663, MS K710,

Los Alamos, NM 87545, USA. Email: dct@lanl.gov.

Abstract

Data mining is a technique for discovering useful information from large databases. This technique is currently being profitably used by a number of industries. A common approach for information discovery is to identify association rules which reveal relationships among different items. In this paper, we use this approach to analyze a large database containing medical-record data. Our aim is to obtain association rules indicating relationships between procedures performed on a patient and the reported diagnoses. Random sampling was used to obtain these association rules. After reviewing the basic concepts associated with data mining, we discuss our approach for identifying association rules and report on the rules generated.

Key Words: Data Mining, Association Rules, Procedure Code, Diagnosis Code.

1 Introduction

Data Mining is a technique for discovering useful information from large databases. A large database represents a huge amount of information which can be potentially very useful if extracted and summarized correctly. Using statistical tools and modeling techniques, one can discover interesting and hidden patterns in the data. These patterns may not be easily detected using traditional methods. Data mining technology is currently being used by a number of industries–including chemical, financial, pharmaceutical and insurance [5, 10, 12, 13].

There are two main reasons for this trend. First, the decreasing cost of electronic storage media has made it economically feasible for businesses to collect and maintain large, long-term databases. Second, analyzing the data and extracting useful information can be potentially very profitable to a business.

For example, reference [12] discusses how data mining techniques were useful in increasing profits of a drugstore in Japan. That study shows how the discovery of high association between sanitary goods and pain relievers helped in increasing the sale of pain relievers by 50%. These pain relievers, which are now sold at regular price, were previously being sold at a discount. Strong associations such as “People with children tend to buy life insurance policies more often than others” and “Owners of sports utility vehicles are more likely to have wireless phones” can be extremely useful for target marketing. Industries can focus their marketing efforts only on a subset of consumers who are very likely to be able to afford and consume their products. This will reduce expenses and increase sales and revenue.

The data mining technique draws ideas from a number of disciplines–including statistics, machine learning and database systems. Ideas from these disciplines are used to characterize the useful information which may be extracted from a large database and to quantify the usefulness of this information. References [5, 10, 13] provide extensive discussions on this topic.

Supermarkets are often used to illustrate the potential benefits of data mining [1, 5, 10, 13, 19]. Suppose a supermarket records the set of items bought by each customer. This data is commonly referred to as “market-basket data”. By analyzing this data, the supermarket may be able to identify which subsets of items are likely to be bought together by a customer. Using this information, the supermarket could, for example, arrange items which are likely to be bought together in adjacent shelves so that it becomes easy for a customer to shop in the store. We will use the supermarket example to review some basic concepts in data mining in Section 2. References [7, 9, 10, 13] provide examples of other types of useful information that may be extracted from large commercial databases.

The focus of this paper is on the application of data mining to medical-record data. The raw data[1] consists of millions of “claims” for medical procedures. Each patient may have several claims, and each claim may have several line items or records, one for each medical procedure performed. A diagnosis was reported with each procedure. Our main goal is to discover relationships between procedures performed on a patient and the reported diagnoses. One common method for discovering such relationships is to look for **association rules** between procedures and diagnoses. (A brief discussion on the concept of association rules is provided in Section 2.) Discovery of association rules is an important component of data mining.[2] Association rules have been widely used by the retail industry under the name “market-basket analysis”. However, the concept of association rules is general and has wide applicability. The purpose of this paper is to demonstrate its applicability to medical data.

The remainder of this paper is organized as follows. Section 2 provides a brief review of association rules and related concepts. In Section 3 we discuss the approach used to obtain association rules from the medical data. Section 4 presents a discussion of some of the rules discovered by our approach. Conclusions and directions for future work are provided in Section 5.

[1] Due to the sensitive nature of the data, we are unable to disclose its source.

[2] Other data mining methods and tools include clustering, classification tree, discriminant analysis, regression analysis, neural networks, genetic algorithms, random sampling, validation etc.

No.	Items Purchased
1	{Bread, Butter, Cereal, Juice, Milk}
2	{Cereal, Juice, Milk}
3	{Bagels, Butter, Cereal, Juice, Milk}
4	{Bread, Cereal, Jelly, Juice, Milk}
5	{Bagels, Jelly, Juice, Milk}
6	{Jelly, Juice, Milk}

Table 1: A small set of transactions

2 Basic Concepts

To explain the basic concepts of data mining, we recall the example of supermarket data. In Section 3, we discuss the similarities and the differences between medical-record data and market-basket data.

Consider a small store that sells the following set of items: {Bagels, Bread, Butter, Cereal, Jelly, Juice, Milk}. List of items bought by six hypothetical customers are shown in Table 1. This table will be used to illustrate the concepts presented in this section. Each row of the table is referred to as a **transaction**.

Definition 1:

1. Given a set S of items, any nonempty subset of S is called an **itemset**.

2. Given an itemset I and a set T of transactions, the **support** of I with respect to T, denoted by support$_T(I)$, is the number of transactions in T that contain all the items in I.

3. Given an itemset I, a set T of transactions and a positive integer α, I is a **large itemset** with respect to T and α if support$_T(I) \geq \alpha$. We refer to α as the **support threshold**.

For simplicity, when the transaction set T is clear from the context, we use “support” instead of “support with respect to T”. The following example illustrates the concepts presented in Definition 1.

Example 1: Let S denote the set {Bagels, Bread, Butter, Cereal, Jelly, Juice, Milk} and let T denote the set of transactions shown in Table 1.

Examples of itemsets are $I_1 = \{\texttt{Cereal}, \texttt{Juice}, \texttt{Milk}\}$ and $I_2 = \{\texttt{Bagels}, \texttt{Bread}\}$.

The support of itemset I_1 with respect to T is 4 since I_1 appears in exactly four of the transactions shown in Table 1. The support of itemset I_2 with respect to T is zero since no transaction in T contains both `Bagels` and `Bread`.

If we set the support threshold at 3, then I_1 is a large itemset since the support of I_1 is 4. Another large itemset for this support threshold is $I_3 = \{\texttt{Jelly}, \texttt{Juice}, \texttt{Milk}\}$, which has a support of 3. Since the support of I_2 is zero, I_2 is not a large itemset for any positive support threshold. ■

Definition 2:

1. An **association rule** is a pair of disjoint itemsets. If L and R denote the two disjoint itemsets, the association rule is written as $L \Longrightarrow R$.

2. The **support** of the association rule $L \Longrightarrow R$ with respect to a transaction set T is the support of the itemset $L \cup R$ with respect to T.

3. The **confidence** of the rule $L \Longrightarrow R$ with respect to a transaction set T is the ratio support($L \cup R$)/support(L).

We once again use the transaction set T defined in Table 1 to illustrate the concepts presented in Definition 2.

Example 2: Consider the itemsets $A_1 = \{\texttt{Juice}, \texttt{Milk}\}$ and $A_2 = \{\texttt{Cereal}\}$. Since A_1 and A_2 are disjoint, $A_1 \Longrightarrow A_2$ (or equivalently, $\{\texttt{Juice}, \texttt{Milk}\} \Longrightarrow \{\texttt{Cereal}\}$) is an association rule. Let R_1 denote this association rule.

The support of R_1 is the support of the itemset $\{\texttt{Juice}, \texttt{Milk}, \texttt{Cereal}\}$. From Table 1, it can be seen that this support value is 4.

Also from Table 1, the support of the itemset $\{\texttt{Juice}, \texttt{Milk}\}$ is 6. Therefore, the confidence of Rule R_1 is $4/6$ or 66.67%. ■

A given association rule $L \Longrightarrow R$ is deemed significant if it has high support and high confidence. In the context of a supermarket, such a rule indicates that a customer who buys the items in set L is also likely to buy the items in set R. For a given support threshold, rules with larger confidence values are more significant than those with smaller confidence values. From the definition of confidence, it can be seen that the confidence value is at most 1. So, sometimes (as in Example 2) confidence is measured as a percentage, rather than a ratio. We are now ready to discuss how these concepts may be applied to medical data.

3 Generating Association Rules from Medical Data

3.1 Structure of Raw Data

The database which was available for our research consisted of over 100 million line items generated over a period of one year. However, due to computational and storage limitations, we analyzed only the first 30 million line items of this database. Each line item has three fields, namely a patient identification code, a procedure code and a diagnosis code. All three fields of each line item are stored as strings of characters. The fields are separated by tabs. Each patient is assigned a unique identification code. Also, different procedures (diagnoses) have different codes associated with them. For each procedure code, a description of the corresponding medical procedure is given in [14]. Similarly, the medical diagnosis corresponding to a diagnosis code is given in [8].

3.2 Interpreting Medical Data as Market Basket Data

We employed a preprocessing step to convert the raw data into a form similar to market basket data discussed in Section 2. From the 30 million line items, we generated for each patient, the set of all procedures performed and all the diagnoses reported. The resulting data set had procedure and diagnosis information for 1,257,645 patients. There were 7,365 different procedure codes and 9,383 different diagnosis codes. The correspondence between the resulting data and the market basket data discussed in Section 2 is as follows. The set of all procedure and diagnosis codes corresponds to the set of items. Thus, the set of all items has $9,383 + 7,365 = 16,748$ elements. The set of procedure and diagnosis codes for each patient is taken to constitute one transaction. In other words, the preprocessing step produced a collection of 1,257,645 transactions. Each itemset consists of one or more procedure/diagnosis codes. Thus, the support of an itemset I is the number of patients (transactions) whose set of items includes all the items in I. Using α to denote the support threshold for a large itemset, an itemset I is a large itemset if at least α patients have all the procedure and diagnosis codes included in I.

The main goal of our study was to discover relationships between procedures performed on patients and the corresponding diagnoses. As mentioned in Section 2, a natural way to discover such relationships is to find association rules of the form $L \Longrightarrow R$, where L is a set of procedure codes and R is a set of diagnosis codes. For such a rule to be significant, it must have high support and confidence.

3.3 Differences between Medical Data and Market Basket Data

Even though we have tried to represent the medical data in a form similar to market basket data, there are a few significant differences between the two datasets. This section highlights the differences.

In market basket data, all items are pooled together whereas in medical data, the items are divided into two categories, namely procedure codes and diagnosis codes. Unlike market basket data, the association between procedure codes and diagnosis codes reveals cause-effect relationships.

Each transaction in the medical dataset refers to a single patient. This can lead to some unexpected association rules as shown in the following example. Consider the example of a patient with heart disease who is involved in an accident requiring treatment for a broken arm. At the end of the year, when we gather all diagnoses and procedures performed on this patient, we may come up with an unexpected association rule such as $\{\text{Cast}\} \Longrightarrow \{\text{Heart Disease}\}$. Of course, a data mining tool will report the above association rule only when a sufficient number of patients who have heart disease also undergo the procedure of obtaining a cast.

In market basket data, the items in a transaction are used for an entire household. An association between diapers and beer is easy to explain when applied to a household instead of a single consumer. A household may include several adults who drink beer and small children who need diapers. This is in contrast to medical data where all the items are associated with a single patient.

3.4 Obtaining Association Rules

The first step in the generation of association rules is the identification of large itemsets. A number of algorithms have been proposed in the literature for obtaining large itemsets (see [13] for a good discussion on this topic). A commonly used algorithm for this purpose is the *Apriori* algorithm presented in [1, 17]. We now provide a brief description of the algorithm focusing on its essential features. For details regarding how the various steps of the algorithm can be efficiently implemented, we refer the reader to [1, 17].

Algorithm Apriori relies on the following subset principle: Every nonempty subset of a large itemset must itself be a large itemset. The algorithm uses this principle in a bottom-up manner. Assume that a support threshold of α is used. Let L_i denote the collection of large itemsets where each itemset has i items. The algorithm begins by identifying all the sets in L_1. This can be done in a straightforward manner by counting the number of transactions in which each item appears, and retaining only those items that appear in at least α transactions. Each item that has the necessary support forms a singleton large itemset and is included in L_1. By the subset principle, items that appear in less than α transactions cannot be part of any large itemset; therefore, they are dropped from further consideration. The collection L_2 can be constructed by considering each pair of sets in L_1 and retaining only those pairs that appear in at least α transactions. In general, having constructed L_i, the collection L_{i+1} is constructed by considering pairs of sets, one from L_i and another from L_1 and eliminating those for which the support is smaller than α. This procedure is continued until all large itemsets up to the desired maximum size have been obtained.

Since our dataset is large (over 16,000 items and one million transactions), a direct application of the Apriori algorithm would need an extremely large amount of computation time. So, we used the idea of **random sampling** from the literature [15, 19, 20] to reduce the computation time.

The underlying method selects a small random sample from the large collection of transactions and runs the Apriori algorithm on the small sample. In doing so, a smaller support threshold is chosen for deciding whether or not an itemset is a large itemset. Once large itemsets based on the random sample are available, support values of these itemsets with respect to the entire set of transactions can be computed. In our implementation of the Apriori algorithm with the sampling step, we partitioned the transaction set into five disjoint sets each consisting of about 250,000 transactions. From each subset of transactions, we generated random samples of size 5,000. We ran the Apriori algorithm on these samples with a support threshold of 5. To further reduce the computation time, we decided to restrict our attention to large itemsets consisting of seven or fewer items. After generating large itemsets from the sample, we used another computer program to determine the support of the large itemsets with respect to the entire set of transactions. The support values for the resulting large itemsets varied from 56 to 1280.

The large itemsets generated consisted of three types: sets that contain only procedure codes, sets that contain only diagnosis codes and sets that contain both procedure and diagnosis codes. Since the focus of this study was to obtain relationships between procedures and diagnoses, we restricted our attention to large itemsets containing both procedure and diagnosis codes. From each large itemset, we formulated one association rule with all the procedure codes on the left and all the diagnosis codes on the right. For each such association rule, we computed the confidence value with respect to the entire set of transactions. We eliminated rules with confidence less than 65%. Also, some of the rules were subsets of other rules. (This is a consequence of the subset principle stated earlier.) After eliminating such rules, we were left with a collection of 11 association rules shown in Table 2. For each rule, the table indicates the support and confidence. The rules are listed in decreasing order of confidence. A discussion of these rules is provided in the next section.

4 Discussion of Results

Most of the rules in Table 2 indicate a rationalizable correspondence between a set of procedures and a diagnosis. Below we discuss the rules in the order in which they are presented in the table. Rule No. 1 shows that a total of 173 patients had undergone the six procedures, namely Vaginal ultrasound, Surgical pathology, Pregnancy test, Hematology, Induced abortion and Penicillin injection, and were diagnosed with “Legally induced abortion.” The confidence level of 99.42% suggests that for virtually all the patients who had undergone the six procedures, the diagnosis was “Legally induced abortion.”

Rule No. 2 also illustrates a logical relationship between the procedures pertaining to pulmonary tests, inhalation treatments and the diagnosis of “Asthma.” The high confidence value for Rule No. 3 is somewhat perplexing. One can see a reasonable correspondence between Debridement of nails and Dermatophytosis but surgery of intestine, urethra and bladder seem to have no direct relationship with Dermatophytosis. One plausible explanation for such a rule is that perhaps among older patients these problems occur concurrently. In fact, this points out a weakness of the association rule approach. As the approach does not use any knowledge of the underlying domain, not all the rules generated are meaningful. This reminds us that data mining is just a tool that provides businesses with a method of generating hypotheses. It does not verify the hypothesis; nor does it provide any information regarding the value of that hypotheses to the business. These hypotheses must be analyzed and verified by people with domain knowledge and expertise. See [11] for further details.

Rule No. 4 confirms the common experience of pregnant women who are given Antibody screening, Rh $\&$ Blood typing, Hepatitis and Rubella testing, etc. It is possible to provide a rational explanation for Rule No. 5. Patients are diagnosed with “Diabetes” on the basis of glycated and blood

No.	Association Rule	Support	Confidence
1	{Vaginal ultrasound; Surgical pathology; Pregnancy test; Hematology; Induced abortion; Penicillin injection} ⟹ {Legally induced abortion}	173	99.42%
2	{Pulmonary bronchospasm evaluation; Pulmonary vital capacity test; Non-pressurized inhalation treatment for acute airway obstruction; Doctor's office visit } ⟹ {Asthma}	56	91.80%
3	{Debridement of nails, manual, five or less; Debridement of nails, each additional, five or less; Intestine excision: Enteroenterostomy, anastomosis of intestine with or without cutaneous enterostomy; Transurethral surgery (Urethra and bladder)} ⟹ {Dermatophytosis}	619	91.43%
4	{Antibody screen (RBC); Rh typing; Blood typing; Hepatitis B antigen; Rubella test} ⟹ {Normal pregnancy}	1178	84.87%
5	{Radiological examination of chest front and lateral view; Automated multi-channel test: one or two clinical chemistry tests; Glycated chemistry tests} ⟹ {Diabetes mellitus}	320	81.63%
6	{Stomach excision; Pyloroplasty; Change of gastronomy tube; Gastrojejunostomy} ⟹ {Schizophrenic disorder}	931	78.70%
7	{Cardiac catheterization: selective left ventricular or left aterial angiography; Cardiac catheterization: selective coronary angiography; Imaging supervision, interpretation and report for injection procedures during cardiac catheterization, ventricular and/or atrial angiography; pulmonary angiography, aortography, and/or selective coronary angiography including venous bypass grafts and arterial conduits} ⟹ {Other forms of chronic ischemic heart disease}	1280	76.97%
8	{Ultrasonic guidance for amniocentesis, radiological supervision and interpretation; Amniocentesis; Amniotic fluid or chorionic villus cells studies; Chromosome analysis additional karyotypes} ⟹ {Known or suspected fetal abnormality affecting management of mother}	81	73.64%
9	{Cardiac catheterization: selective left ventricular or left aterial angiography; Cardiac catheterization: selective coronary angiography; Imaging supervision, interpretation and report for injection procedures during cardiac catheterization, ventricular and/or atrial angiography; Pulmonary angiography, aortography, and/or selective coronary angiography including venous bypass grafts and arterial conduits} ⟹ {Symptoms involving respiratory system and chest}	1152	69.27%
10	{Radiological examination of chest front view; Generic doctor's office visit; Initial in-patient hospital consultation} ⟹ {Symptoms involving respiratory system and chest}	637	68.86%
11	{Culture, bacterial screening only, for single organisms; Identification of culture, bacterial, urine quantitative colony count; Smear, primary source with interpretation, routine stain for bacteria, fungi or cell types} ⟹ {Inflammatory disease of cervix, vagina or vulva}	133	67.51%

Table 2: Association Rules Obtained

chemistry tests. Also, since such patients are more likely to develop cardiac problems, the radiological examination of the chest seems justified.

However, Rule No. 6 indicates an unclear correspondence between the diagnosis of "Schizophrenia" and stomach related procedures. We believe that this rule was generated because a large fraction of the patients who were diagnosed with both Schizophrenic disorder and stomach disorder shared stomach-related procedures but had no common stomach-related diagnosis.

Rules No. 7 and No. 9 are very similar. The procedures listed in these rules are all reasonable for diagnosing patients with heart, chest and respiratory problems. Rule No. 8 also illustrates a predictable set of of procedures related to fetal abnormality. Rule No. 10 shows that patients who have respiratory or chest problems undergo radiological examination, pay a visit to the doctor and have in-patient hospital consultation. Finally, Rule No. 11 provides a meaningful correspondence between diseases related to cervix, vagina and vulva and procedures such as Culture & bacterial screening, Identification of culture, bacterial, urine quantitative colony count, Smear, etc.

Overall, all the rules except No. 3 and No. 6 indicate a clear correspondence between procedures and diagnoses. The quantitative information included in the rules can be potentially very revealing and beneficial to medical professionals.

The above discussion shows how application of association rules to medical data may be of interest to physicians. Similar analysis tools can be applied to detect fraudulent behavior among physicians and patients. Association between a group of physicians and patients may be indicative of a *ping-ponging* scheme.[3] Association of a physician with frequent prescription of expensive medication and non-invasive procedures (such as aroma-therapy) may be of interest to the administrators of government sponsored medical programs and insurance companies. Such situations may be indicative of other types of fraudulent behavior (e.g. the physicians receiving kickbacks). In this paper, we focused our attention on finding associations between procedure and diagnosis codes. A physician

[3] A scheme in which a group of doctors refer their patients back and forth to increase their pay-offs.

who is new to the field may benefit substantially by knowing the set of commonly performed procedures for a particular diagnosis. Association rules can also provide an indication of collections of diagnoses that are correlated and are likely to occur together.

5 Conclusions and Future Work

We discussed the application of a standard data mining technique to the case of medical data. We obtained association rules that show relationships between medical procedures and the corresponding diagnoses. The association rules provide a method of measuring joint frequencies for common combinations of medical procedures and the corresponding diagnoses.

Our results warrant further application and development of data mining techniques for medical informatics. A direct extension of our work is to construct association rules among procedures and among diagnoses. For example, an association rule of the form $L \Longrightarrow R$ where both L and R are subsets of procedures, provides quantitative information regarding how often the set of procedures in R are performed on patients given that the set of procedures in L is performed. Another way of extending our study is to use some knowledge of the underlying domain to decide the types of association rules that would be beneficial to medical professionals. Finally, it may be useful to carry out data mining on a restricted data set that includes only a specific class of diseases and the relevant procedures and treatments. Such a study may provide useful information concerning the effectiveness of a set of procedures for diagnosing a particular disease in the class or that of a set of treatments in curing a particular disease.

References

[1] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, *Proc. 1993 ACM SIGMOD*, Washington, DC, May 1993, pp. 207–216.

[2] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules”, *Proc. 20th International Conference on Very Large Databases* (VLDB’94), Santiago, Chile, Sept. 1994, pp. 487–499.

[3] R. Agrawal and J. C. Schafer, “Parallel Mining of Association Rules”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, 1996, pp. 962–969.

[4] C. C. Aggarwal and P. S. Yu, “Mining Large Itemsets for Association Rules”, *Bulletin of the Technical Committee on Data Engineering*, Special Issue on Mining of Large Databases, Vol. 21, No. 1, March 1998, pp. 23–31.

[5] D. Barbará (Editor), *Bulletin of the Technical Committee on Data Engineering*, Special Issue on Mining of Large Databases, Vol. 21, No. 1, March 1998.

[6] S. Brin and L. Page, “Dynamic Data Mining: Exploring Large Rule Spaces by Sampling”, Manuscript, Department of Computer Science, Stanford University, Stanford, CA, 1998.

[7] C. Bettini, X. S. Wang and S. Jajodia, “Mining Temporal Relationships with Multiple Granularities in Time Sequences”, *Bulletin of the Technical Committee on Data Engineering*, Special Issue on Mining of Large Databases, Vol. 21, No. 1, March 1998, pp. 32–38.

[8] B. Christmas, B. Ericson, T. Hodges, et al., *Medicode’s International Classification of Diseases* (ICD-9-CM), Fifth Edition, Medicode Publications, Salt Lake City, UT, 1995.

[9] D. J. Cook, L. B. Holder and S. Djoko, “Knowledge Discovery from Structural Data”, CESDIS Technical Report Series No. TR-95-149, Goddard Space Flight Center, Greenbelt, MD, 1995.

[10] K. M. Decker and S. Focardi, “Technology Overview: A Report on Data Mining”, Technical Report CSCS TR-95-02, Swiss Scientific Computing Center, 1995.

[11] H. A. Edelstein, “Introduction to Data Mining and Knowledge Discovery”, Two Crows Corporation, 1997.

[12] Y. Hamuro, N. Katoh, Y. Matsuda and K. Yada, “Mining Pharmacy Data Helps to Make Profits”, *Data Mining and Knowledge Discovery*, Vol. 2, 1998, pp. 391–398.

[13] M. Holsheimer and A. P. J. M. Siebes, “Data Mining: The Search for Knowledge in Databases”, Technical Report CS-R9406, CWI, Amsterdam, The Netherlands, 1994.

[14] C. G. Kirschner, L. M. Frankel, J. A. Jackson, et al., *Physician’s Current Procedural Terminology* (CPT’96), American Medical Association, Chicago, IL, 1996.

[15] S. D. Lee, D. W. Cheung and K. Ben, “Is Sampling Useful in Data Mining? A Case in the Maintenance of Discovered Association Rules”, *Data Mining and Knowledge Discovery*, Vol. 2, No. 3, Sept. 1998, pp. 233–262.

[16] N. Pasquier, Y. Bastide, R. Taquil and L. Lakhal, “Efficient Mining of Association Rules Using Closed Itemset Lattices”, *Information Systems*, Vol. 24, No. 1, 1999, pp. 25–46.

[17] R. Srikant, “Fast Algorithms for Mining Association Rules and Sequential Patterns,” Ph.D. Thesis, Department of Computer Science, University of Wisconsin, Madison, WI, 1996.

[18] R. Srikant and R. Agrawal, “Mining Generalized Association Rules,” *Proc. 21st International Conference on Very Large Databases* (VLDB’95), Zurich, Switzerland, Sept. 1995, pp. 407–419.

[19] H. Toivonen, “Discovery of Frequent Patterns in Large Data Collections”, Ph.D. Thesis, Report A-1996-5, Department of Computer Science, University of Helsinki, Finland, 1996.

[20] H. Toivonen, “Sampling Large Databases for Association Rules”, *Proc. 22nd International Conference on Very Large Databases* (VLDB’96), Mumbai, India, Sept. 1996, pp. 134–145.